USING A MULTIMODAL IMMERSIVE ENVIRONMENT TO INVESTIGATE PERCEPTIONS IN AUGMENTED VIRTUAL REALITY SYSTEMS

Samuel Chabot, Wendy Lee, Rebecca Elder, and Jonas Braasch

Graduate Program in Architectural Acoustics
Rensselaer Polytechnic Institute
Greene Building, 110 8th St,
Troy, NY 12180, USA
chabos2@rpi.edu

ABSTRACT

The Collaborative-Research Augmented Immersive Virtual Environment Laboratory at Rensselaer is a state-of-the-art space that offers users the capabilities of multimodality and immersion. Realistic and abstract sets of data can be explored in a variety of ways, even in large group settings. This paper discusses the motivations of the immersive experience and the advantages over smaller scale and single-modality expressions of data. One experiment focuss on the influence of immersion on perceptions of architectural renderings. Its findings suggest disparities between participants' judgment when viewing either two-dimensional printouts or the immersive CRAIVE-Lab screen. The advantages of multimodality are discussed in an experiment concerning abstract data exploration. Various auditory cues for aiding in visual data extraction were tested for their affects on participants' speed and accuracy of information extraction. Finally, artificially generated auralizations are paired with recreations of realistic spaces to analyze the influences of immersive visuals on the perceptions of sound fields. One utilized method for creating these sound fields is a geometric ray-tracing model, which calculates the auditory streams of each individual loudspeaker in the lab to create a cohesive sound field representation of the visual space.

1. INTRODUCTION

As sets of data become ever larger and more complex, tools and practices for extracting information must also adapt and develop. One such emerging tool is that of the immersive virtual environment. These state-of-the-art spaces present users with enveloping visuals often combined with three-dimensional audio reproduction systems. Size and degree of immersion in the visual or auditory field can vary, ranging from the incredibly large and nearly all-encompassing to the more individual and limited scale. The AlloSphere [1] at The University of California, Santa Barbara, would be an example of the former, supporting upwards of 30 simultaneous on a bridge spanning the center of a spherical projection screen. Currently, 54 loudspeakers form three rings around the sphere for spatial audio using vector-based amplitude panning (VBAP) and Ambisonics techniques [2]. Many of the CAVE (Cave Automatic Virtual Environment) systems reside in the latter category and are generally more appropriate for a single user or two [3]. Audio playback is often done simply through a 5.1 surround sound system.

An immersive workspace at Rensselaer Polytechnic Institute has been developed to facilitate large groups of users while not being rendered intractably complex: The Collaborative-Research Augmented Immersive Virtual Environment Laboratory, or CRAIVE-Lab. A function of this lab is its multimodality which puts equivalent emphasis on both the auditory display and visual display. This workspace measures $10\,\mathrm{m}$ by $12\,\mathrm{m}$ to accommodate groups of users both small and large. A nearly 360° projector screen surrounds the perimeter of the space and reaches $4.3\,\mathrm{m}$ tall. The screen is rectangular in shape, with rounded corners between the four sides [r=1.52m]. A sole PC equipped with the software Pixelwarp Evo[1] warps and blends eight overlapping high-performance projectors to produce a single continuous desktop display with a final resolution of 15360 by 1200 pixels. Therefore, content-creation is a rather straight-forward process. Visuals need only be created to encompass a screen width of 15k pixels. For many programs and resources, such as Photoshop or HTML, this only requires defining the workspace with the correct pixel dimensions. Rapid prototyping of immersive recreations of real environments and presentations of abstract and symbolic data is therefore possible. Figure 1 and 3 show respective examples of this. The projection screen itself is made of a microperforated PVC material, allowing it to remain acoustically transparent for the horizontal array of 128 loudspeakers behind it.

The array is positioned at approximately ear-height and follows the perimeter of the lab. Six additional loudspeakers are hung from the ceiling and directed downward into the workspace to add a third dimension to the auditory display. Matching the height and length of the projector screen, a heavy acoustic curtain hangs behind the horizontal array to provide strong damping of the physical room's response. Carpeting has been installed over the previously concrete floor to provide additional dampening of sound and reduce reflections of light. The entire loudspeaker array is controlled using a single Mac Pro equipped with an RME HDSPe MADI FX sound card[2]. This allows channels to be accessed and addressed individually. Production of sound is therefore extremely flexible, and multiple approaches can be taken. For example, individual

[1] http://pixelwix.com/warp-and-blend-software/10-pixelwarp-evo-standard-wide-screen-warper.html

[2] http://www.rme-audio.de/en/products/hdspe_madi_fx.php

Figure 1: Panoramic rendering of realistic environment featuring two users

channels can be explicitly targeted with auditory streams. Due to the high-density nature of the array, the system also allows for complex spatialization techniques such as higher-order ambisonics and wave field synthesis, which require channels to work in conjunction to form a cohesive sound field. For a much more detailed overview of the CRAIVE-Lab and its hardware and software considerations, refer to this previous ICAD paper [4].

A variety of studies, including use-cases, have been performed in the lab. One study demonstrated the advantages of the space's multimodality in an experiment involving abstract data exploration. Participants were presented with a large volume of visual data. Auditory cues and presentation techniques were tested for aiding visual data extraction. A variety of cues were chosen, including sine tones, white noise and camera shutter sounds, and three playback groupings were available: local, regional, or global sounds. Participants were scored on their performance, which was rated for speed and accuracy of identification.

Focusing on the advantages of human-scale immersive presentations of realistic scenes, another study investigated the perceptions of architectural renderings when viewed both on two-dimensional printouts and within the immersive visual display. Participants were presented with three architectural renderings that had been "inserted" into real scenes and asked to rank their preferences (first with the printouts, second using the immersive display). Its findings suggest disparities between participants perceptions of the renders when using either two-dimensional printouts or the immersive CRAIVE-Lab screen.

Another project seeks to study the connection and influences of immersive imagery on users' perceptions of immersive sound fields. Renderings of realistic scenes are generated and paired with acoustic simulations of spaces, produced using panoramic imagery and a ray-tracing model. To do so, a geometric model is implemented using MATLAB to calculate individual impulse responses at each loudspeaker location. These impulse responses are then used to create the auditory streams of each loudspeaker for a combined sound field recreation of the modeled space. Users' perceptions of the artificial sound fields will first be surveyed without any immersive imagery. Later, users will be presented with paired visual scenes (sometimes with correctly matched and sometimes with disparate scenes) to again be assessed on their perceptions of the sound fields amidst the immersive visuals.

2. AUDITORY CUES FOR EXTRACTING VISUAL DATA

The question this psychophysical study sought to address was which auditory cue and cue type would most effectively assist visual data extraction in a complex, immersive visual display system [5]. Previous works in this area both discuss human visual and auditory attention [6] and find auditory information altered the perceptions of visual information [7]. In this study, it was hypothesized that auditory cues that provide the strongest localization information (while also inflicting the least attentional distractions) would provide the most effective assistance for speed and accuracy of data extraction. An objective of the experiment was to assess various auditory cues and presentation techniques to determine those combinations that most helped participants' performances.

2.1. Experimental Design

Seven participants (various graduate students; 4 male, 3 female; ages between 24-33) were recruited for testing. The task presented to each user was to extract visual data from the immersive CRAIVE-Lab screen using auditory cues and answer ques-

Figure 2: Sample of icons used to populate the entire CRAIVE-Lab panoramic screen

tions based on the visual data as quickly and accurately as possible, modeled on the test method described in *The Handbook for Multisensory Processes* as "spatial discrimination tasks" [8]. To do so, the CRAIVE-Lab panoramic screen was populated with icons which depicted a variety of symbols: letters, numbers, animals, foods, etc. Figure 2 depicts a sample. The icons were arranged in a 92-by-9 grid (for a total of 828) on a plain white background across the entirety of the screen. After assessing various layouts, this configuration was chosen for its combination of legibility and density of icons.

For the auditory cues to be tested, ten different sounds were chosen for assessment of their localization and distraction levels [9]. Each cue was repeated over a question's entire 20-second duration (with the exception of one, a camera shutter cue, which only played a single instance). The variety of the ten cues included the following: sine tones at 500 Hz and 1000 Hz, presented as bursts or continuous streams, white noise, also presented as bursts or a continuous stream, a click train, bell ding, tech alert, and camera shutter (former seven self-produced, latter three sourced from Zapsplat[3]). The cues were presented in one of three ways: locally, where cues were presented from a single loudspeaker at the target location; regionally, where cues were presented using an area of 21 loudspeakers surrounding the target location; or globally, where cues were routed directly to all 128 loudspeakers in the lab. Regional presentation utilized the loudspeaker at the target location and the ten loudspeakers to the left and to the right. For reliability, all ten auditory cues were tested using all three presentation methods three times each, for a total of 90 test questions throughout the experiment. Users interacted with the test questions on a separate laptop using an interface created in MATLAB, pictured in figure 4. This interface would present users with the test questions and automatically record the users' speed and accuracy upon answering. The MATLAB interface took advantage of the Open Sound Control (OSC) protocol to appropriately engage each auditory cue for each question. The cues were handled by Max/MSP on the lab's Mac Pro linked to the loudspeaker array. To avoid potential biases, the order of the cues and presentation methods was randomized (and documented) for each user by MATLAB before outputting to Max.

[3]https://www.zapslat.com

2.2. Procedure

A participant's session went as follows: The participant was briefed with instructions on the laptop before entering the lab. The end of the instructions had the participant bring the laptop into the lab, where the visuals were on display, and place it on a (movable) stand in the center of the room. An 8-ft by 8-ft box marked by white tape on the floor indicated the space a participant may move about but was required to remain within. This was to prevent the user from drifting too close to both the screen and loudspeaker array, which would add bias in both the auditory localization and the visual search. Upon clicking "Begin" on the laptop interface, a 3 second "rest" page appeared. These appeared between each question. Then, accompanying the question "What color is the following icon on the big display?", one of the icons was shown with all color removed and a multiple-choice option of four colors. An example of one of these pages is shown in figure 4.

One of the ten cues was presented using one of the three methods, determined by the randomized order. Participants answered the question as quickly and accurately as possible and were instructed not to guess. (It was determined that arbitrary guesses do not inform results of accuracy and speed based on the auditory cues.) Each question was given a maximum of 20 seconds. This was done to prevent users from eventually abandoning the auditory cues and relying solely on visual search. If the question was not answered in time, the cue would stop and the laptop would move to the next "rest" page. No feedback was provided after each question. After each set of ten questions, the participant was allowed an untimed break to prevent fatigue. Upon completing all 90 questions, the participant was asked psychological questions concerning the ten auditory cues. The interface posed three prompts for each cue as it was replayed from a loudspeaker: "Describe the sound in your own words.", "How stressed do you feel after hearing the sound?", and "Do you find the sound pleasant?" Participants did not have a time limit for this section.

2.3. Results and Analysis

Most of the participants correctly identified the sine tones and white noise as such, while the other cues were perceived more ambiguously. For example, participants responded with perceptions of a temple wood block, metronome, and rain drops for the click train. The bell ding and tech alert were most often associated with phone rings, but others described sounds such as bells and water droplets. Participants appeared to draw on familiar conventional sounds to describe the cues. The camera shutter had the greatest consensus, with most responses including the word "camera." The variety in responses reveals potential biases in interpretations, and warrants further research with a larger sample size. Some participants commented on the perceived ability to localize cues. Sine tones were said to offer poor localization information, while more complex cues, such as the bell and shutter, were perceived as localized more easily. The Franssen Effect points out that sine tones lacking a strong onset are particularly difficult to localize by the human ear [10]. White noise bursts and clicks saw the highest accuracy, and both constant sine tones the lowest. The constant white noise and clicks provided the quickest responses, and the bell ding and 500 Hz tones the slowest. Figure 5 shows the mean accuracy of each auditory cue sorted by presentation type. Each bar is an average of the total 21 responses (3 responses each of the 7 participants). Figure 6 shows the mean speed of the same pairings. This second graph considers specifically those questions answered

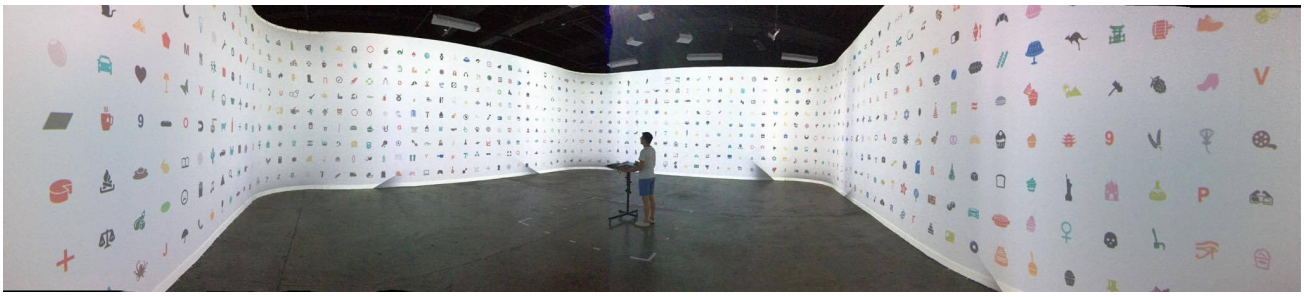

Figure 3: Participant in the CRAIVE-Lab with the final panoramic image with all 828 icons aligned in a 92-by-9 grid layout

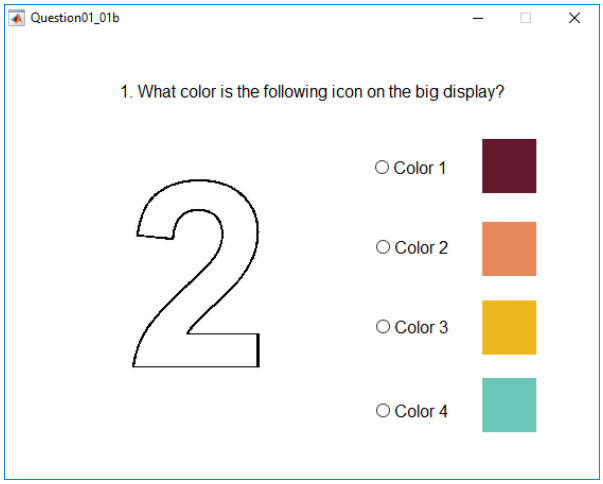

Figure 4: Example of MATLAB interface presenting questions to participant

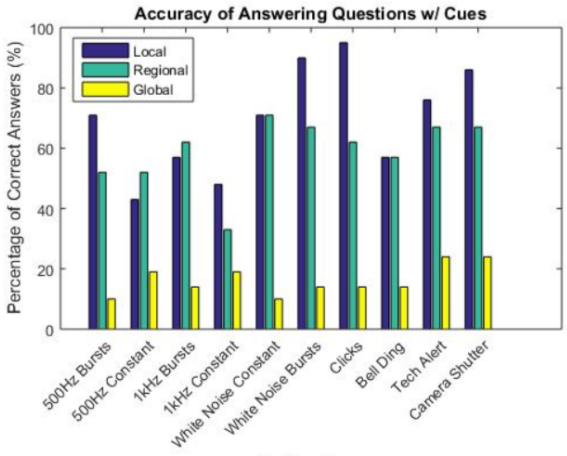

Figure 5: Accuracy of responses of each cue type, shown for each of the three presentation methods (local, regional, global). Each bar is the mean of the total 21 responses (3 responses per participant).

correctly, as questions not answered were counted as incorrect and are included in the accuracy tabulation.

Cues presented locally most often led to both the quickest and most accurate responses. Regional cues provided nearly the same accuracy for some and dramatic drops for others. For the tones, all cues not presented locally saw spikes in identification times, but the more complex tones did not see as large a shift until presented globally. The presentation of cues globally saw a drastic reduction in accuracy and spikes in response times. Cues presented locally provided the best performance, however regional cues also did relatively well. This may indicate that the auditory cues were most useful in informing participants of the general area of the visual target, but that vision took over for the searching technique once the target was regionalized. Unexpected was the high performance elicited from the camera shutter cue. Its accuracy exceeded nearly all other cues for each respective presentation type, and speed of identification did not see a dramatic difference from the other complex tones. It is possible that the single cue was less distracting while still providing adequate localizing information. When asked how stressed a user felt after hearing each sound, only the four more complex sounds (click, bell, tech alert, shutter) averaged below a 3 on a 5-point scale. Five of the six tones and noise scored an average of 3.5 or above, with both constant sine tones reaching a score of approximately 4.8. In response to a cue's pleasantness, the four complex cues received the most favorable answers.

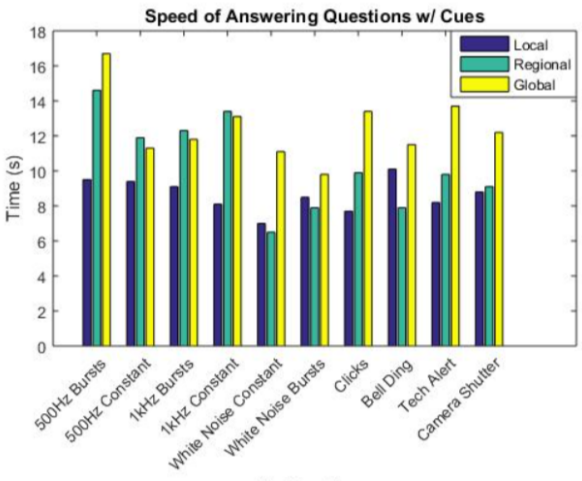

Figure 6: Mean time for correctly answered questions of each cue type, shown for each of the of three presentation methods (local, regional, global).

3. IMMERSIVE ARCHITECTURAL RENDERINGS

Another investigation into the advantages of an immersive system encompassing a congruent audio and visual experience explored questions of architectural design and how it can be better analyzed when presented in a physically immersive space. Through the spaces immersion and multimodality, it is thought that designs can be interpreted at greater depth and more true to reality [11]. Currently, many renderings suffer one or more of the following characteristics: glorified aesthetic; unattainable vantage points; unrealistic environments; no reference to senses other than the visual. Buildings are most often rendered from a birds-eye view, a perspective most clients will never see in reality, along with other ideal conditions, such as perfect lighting, coloring, and atmospheres. Another consequence of most architectural renderings is the two-dimensional presentation of displays or printouts. Clients are then asked to imagine the finished product at full human scale. An immersive augmented environment is the closest form for presentation to human scale.

3.1. Site Data

For this experiment, sites of urban landscapes were reviewed for their diversity of design techniques and implementations. These sites were also reviewed for the activity of their soundscapes. Characteristics desired of each site included public access, close proximity to transportation hubs, and surrounding foliage. The three sites chosen were all located in Manhattan: the plaza at the southeast corner of Central Park at the William Tecumseh Monument, Foley Square, and Washington Square Park. Collection of the site data consisted of two parts. One was the retrieval of the visual footprint. This entailed taking a number of photographs from the same location and ensuring enough photographs were taken to cover the entire 360 degree of the horizontal plane. The camera was positioned at eye-level in order to recreate the most realistic perspective of the scene (versus a bird's eye view, etc.). Post-processing on these images in Photoshop stitched and edited them together to form a cohesive 360 degree image for presentation on the CRAIVE-Lab screen. The other part of site data collection focused on the soundscape. This was done using two Zoom H2s each offset by 45 degrees to capture eight surrounding channels to make use of the CRAIVE-Lab speaker array. To present the most cohesive representation possible, visuals and audio were captured during the same site visit. Architectural renderings were sourced from available student designs within the School of Architecture. Three designs were chosen, ranging from simple to the more abstract and unfamiliar. These three architectural renderings were then "placed" in the visual scenes taken at each site in Manhattan. A location within the image was chosen for the new replacement building. Using Rhino, the architectural software, and Photoshop, the renderings were "inserted" into the panoramic scenes so as to realistically. Figure 7 shows the three renderings within the Site 3 location.

3.2. Testing

The experiment consisted of three tests: the first involved only auditory stimuli; the second involved only a printout of the visual location; and the third combined the audio with the panoramic imagery for the complete immersive experience. A total of 15 participants were tested (graduate students with and without architectural training; ages between 24-33). In Test 1, participants were

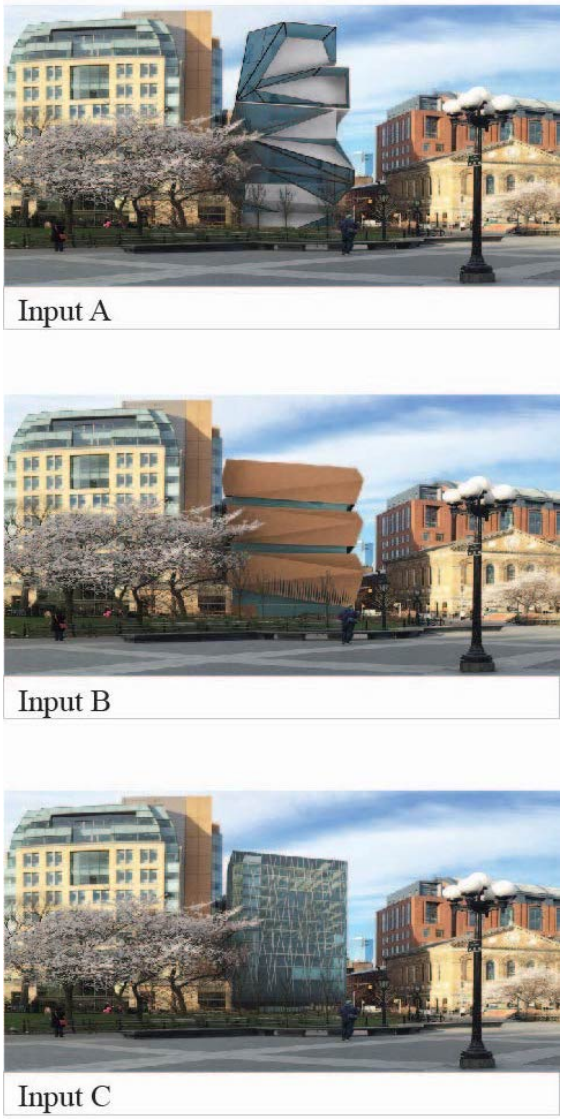

Figure 7: Three renderings placed into Site 3 to be evaluated for their perceived appropriateness in the location

asked to stand in the center of the CRAIVE-Lab. The soundscapes recorded at each of the three sites, lasting two minutes each, were played. This was done prior to exposing the participant to the visual data to prevent individuals from becoming biased toward the visual data while answering perceptual questions about the audio data, such as identifying 3-5 specific sources in the soundscape (e.g. car horn, mechanical noise, etc.). Participants also ranked the soundscapes on a scale of 1 to 5 for 17 pairs of attributes (e.g. wide/narrow, far/nearby, etc.). Test 2 surveyed participants' visual perceptions of each design at each site using a two-dimensional printout. It asked participants to rank the designs by how well they fit in with the surrounding environment. The third test combined the soundscapes previously heard with the panoramic imagery of each site. This provided the participants with much more context about each location and how each rendering fit into the scene. Participants were again asked to identify sound sources and describe

the soundscape, as well as choose the renderings they thought most appropriate for the scene and provide feedback.

3.3. Results and Analysis

Participants' subjective answers for sound identification and ranking may have been altered with the introduction of visuals. Some identifications were more precise when a visual aid was available, such as multiple incorrect identifications of keys rattling in Test 1, but the subsequent correct identifications of dog tags in Test 3 (where a dog was visible in the scene). For Site 1, the identification of people talking as one of the predominant sources during Test 1 rose from 60% of participants to 86% during Test 3 (where multiple groups of people are clearly visible in the foreground). Sites 1 and 3 also saw more consensus in the subjective ranking: the number of pairs where there was over 50% agreement on the score rose from 2 to 8 (of the total 17) and and 5 to 11 for Sites 1 and 3, respectively.

The results of this experiment suggest that the presentation style of architectural renderings can play a role in the perceptions of that rendering. The clearest example of this occurred for the scene of Washington Square Park. During Test 2 where participants only had the two-dimensional printout, replacement building input B was chosen by 33% of participants. However, during Test 3 in the immersive setting, this share shot up to 73%. Participants cited the greater level of context when surveying the scene as having swayed their decision. Two additional buildings around Washington Park with warmer color profiles (like that of replacement building input B) are introduced in the panoramic image versus the two-dimensional printout. Overall, 70% of participants declared the introduction of bi-modal sensory stimuli as influential on their judgments. Figure 8 shows the overall results for each site during both Test 2 and Test 3. It can be seen that responses during the third test do not always match up with those in the second test.

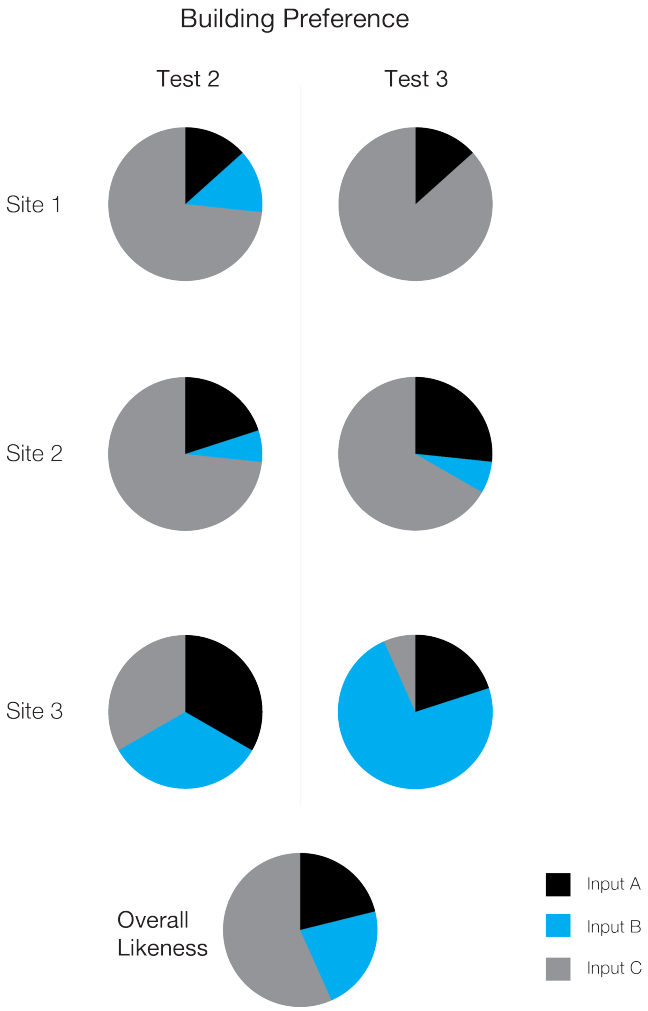

Figure 8: Overall preferences of participants broken down by site and rendering during Test 2 and Test 3

4. INFLUENCES ON USERS' PERCEPTIONS OF SOUND FIELDS BY IMMERSIVE VISUAL IMAGERY

This project is the expansion of the soundscape analysis in the immersive rendering work. It seeks to study the connections and influences of immersive imagery on users' perceptions of immersive sound fields. Namely, how are one's perceptions of a spatialized acoustic environment altered in the presence of human-scale visual environments? A previous study by Yuko Wani [12] presented users with imagery on a single projection screen and played a stereo presentation of an auralization of the space pictured in the imagery. This experiment would build on the CRAIVE-Lab's immersive, multimodal features by presenting users with various immersive sound fields, both with and without paired immersive visuals. In order to investigate this, the sound fields of a variety of spaces must be created and paired with (matching and disparate) panoramic imagery.

4.1. Procedure

To do so, a ray-tracing model was implemented to generate impulse responses for the individual loudspeakers of the Lab. A geometrical model of a space is defined in MATLAB. Figure 9, an example of Cologne Cathedral in Germany, shows sound-reflecting walls (black), a sound source (red), and a receiver (blue). A set of rays is sent out from the sound source in every direction within the horizontal plane at equiangular distances of 5°. Each ray is then traced, and every time a ray meets a wall it is reflected back using Snell's law considering that the outgoing angle equals the incoming angle. The ray is traced until the 20^{th} reflection occurs, unless the ray exits the geometrical model. At every reflection, the sound level is attenuated by 2 dB across frequency to simulate acoustic wall absorption. The sound intensity is also attenuated over distance based on the inverse square law, assuming the sound source to be of omnidirectional character. The collection of rays is shown in figure 9 as gray lines such that the rays become lighter in color with distance and decreasing sound pressure. All rays are then collected at the receiver position assuming a spatial window matching the dimensions of the CRAIVE-Lab. Each calculated ray is tested if it intersects the spatial window at the receiver position. For each intersecting ray, it is then calculated how far it traveled from the source position to the receiver position, at which azimuth angle it arrives at the receiver position and how many times it had been reflected (reflection order).

Based on these data a room impulse response is calculated in which each loudspeaker contributes individually to the entire

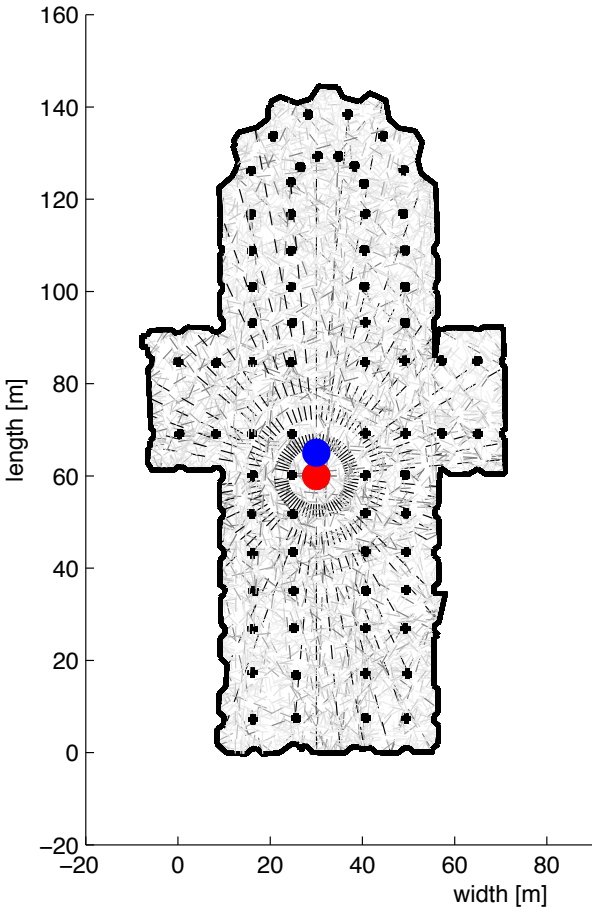

Figure 9: Schematic of a recreated space (Cologne Cathedral) used in the geometric ray-tracing model. Red dot is source location, blue dot is receiver location

sound field those reflections that crossed the threshold of the spatial window at the loudspeaker location. Each impulse contains the rays calibrated to the amplitude each should have based on distance traveled and number of wall reflections. In addition, a late reverberation tail is generated at a constant level (assuming a statistically evenly distributed diffuse reverberation field) using an exponentially decaying Gaussian noise burst.

Using these generated data, one can convolve the impulse responses of each loudspeaker with anechoic tracks to create a sound field that users are able to explore and walk around while maintaining the perceived appropriate source image location.

4.2. Tentative Experimental Setup

The experiment will follow a similar flow as that in the immersive architectural renderings experiment. Participants are to be surveyed on the sound fields and panoramic visuals separately before they are paired together. First, participants will be presented with the generated sound fields to determine their unbiased (by the visuals) perceptions of what (or, rather, where) they are hearing. Various questions will be asked in a thorough questionnaire to determine each sound field's perceived acoustic scene. Second, participants will be shown various panoramic scenes and asked to describe the spaces and their expectations about each acoustic scene. Later, participants will be brought back to experience combinations of sound fields and panoramic scenes. Some of the sound fields will be paired with their corresponding scenes while others will be intentionally mismatched. The questioning for this is twofold: to what extent must an acoustic scene match its visual environment, and can participants be convinced of mismatched sound fields that do not belong to the immersive visuals they are presented with?

5. FUTURE WORK

In the near future, the full experimental setup for analyzing the influence on users' perceptions of immersive sound fields with or without the presence of immersive visuals will be complete and ready for participants. More data across all three studies will be collected as more users participate to better identify patterns in trends in user perceptions. Continuing work seeks to explore the consequences of the congruent immersive visual and immersive audio systems and the perceptual effects that occur when users are presented with both systems simultaneously. Another discussed project would pair panoramic imagery with a real-time user-adjustable acoustic room model. Users would then be instructed to adjust the acoustic profile of the room to match what they perceive to be the most correct profile. Responses would be compared with other users to determine an overall consensus or lack thereof.

6. CONCLUSION

The CRAIVE-Lab fits into the field as a resource for presenting both real and symbolic data, capable of multimodality. The advantages of having simultaneous immersive visual and auditory displays working in conjunction are being studied as qualities unique to environments such as this. These preliminary studies hint at strong relationships between the modalities. Ongoing research will continue to analyze the effects and consequences of such a system on human-computer interactions.

7. ACKNOWLEDGMENT

This work is supported by the National Science Foundation (NSF #1229391) and the Cognitive and Immersive Systems Laboratory (CISL) at Rensselaer directed by Hui Su.

8. REFERENCES

[1] X. Amatriain, J. A. Kuchera-Morin, T. Höllerer, and S. Travis Pope, "The AlloSphere: Immersive multimedia for scientific discovery and artistic exploration," *IEEE Multimedia*, vol. 16, no. 2, pp. 64–75, 2009.

[2] A. Cabrera, J. Kuchera-Morin, and C. Roads, "The Evolution of Spatial Audio in the AlloSphere," *Computer Music Journal*, vol. 40, no. 4, pp. 47–61, 2016.

[3] S. Manjrekar, S. Sandilya, D. Bhosale, S. Kanchi, A. Pitkar, and M. Gondhalekar, "CAVE: An emerging immersive technology -a review," *Proceedings - UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, UKSim 2014*, pp. 131–136, 2014.

[4] S. Chabot and J. Braasch, "An Immersive Virtual Environment for Congruent Audio-Visual Spatialized Data Sonifications," in *Proceedings of the 23rd International Conference on Auditory Display*, Happy Valley, PA, 2017.

[5] W. Lee, "Using Auditory Cues to Perceptually Extract Visual Data in Collaborative, Immersive Big-Data Display Systems," M.S. thesis, Sch. of Arch., Rensselaer Polytech. Inst., Troy, NY, 2017.

[6] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends in Cognitive Sciences*, vol. 12, no. 5, pp. 182–186, 2008.

[7] B. Cowan, D. Rojas, B. Kapralos, K. Collins, and A. Dubrowski, "Spatial sound and its effect on visual quality perception and task performance within a virtual environment," *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, p. 050126, 2013.

[8] G. A. Calvert, C. Spence, and B. E. Stein, Eds., *The Handbook on Multisensory Processes*. Cambridge, Mass.: MIT Press, 2004.

[9] J. M. Terhune, "Localization of pure tones and click trains by untrained humans," *Scandinavian Audiology*, vol. 14, no. 3, pp. 125–131, 1985.

[10] W. M. Hartmann and B. Rakerd, "Localization of sound in rooms IV: The Franssen effect," *The Journal of the Acoustical Society of America*, vol. 86, no. 4, pp. 1366–1373, 1989.

[11] R. Elder, "Influence of Immersive Human Scale Architectural Representation," M.S. thesis, Sch. of Arch., Rensselaer Polytech. Inst., Troy, NY, 2017.

[12] Y. Wani, T. Terashima, and Y. Tokunaga, "Effect of visual information on subjective impression for sound field in architectural space," vol. 19, no. 1, p. 040102, 2013.